AWRA
Community, Conversation, Connections

July 19, 2002

Ganapati P. Patil
Center for Statistical Ecology and Environmental Statistics
Penn State University
421 Thomas Bldg
University Park, PA 16802

Dear Ganapati,

On behalf of the American Water Resources Association Board of Directors and members, it is my pleasure to inform you that you have been selected to receive the 2002 Boggess Award for your paper "Predictability of Surface Water Pollution Loading in Pennsylvania using Watershed-Based Landscape Measurements" which was published in the August 2001 issue of the *Journal of the American Water Resources Association.*

The Boggess Award was established by AWRA in 1973 to honor the author of the best paper published in the *Journal of the American Water Resources Association* during the previous year. The award was established to honor William R. "Randy" Boggess, a charter member of AWRA, one of the first directors, a past president of the association, a former editor and an individual who has made significant contributions to the *Journal of the American Water Resources Association.*

The award will be presented at the Annual AWRA Awards Luncheon Wednesday, November 6, 2002. The luncheon will be held in conjunction with our Annual Water Resources Conference, which will be at the Wyndham Franklin Plaza Hotel in Philadelphia, Pennsylvania. I hope you will attend the luncheon and receive this award.

I would appreciate your sending a one paragraph biographical sketch and a 3"x5" black and white head and shoulders photograph to Kenneth D. Reid, CAE, Executive Vice President of AWRA at the address listed below no later than October 1, 2002. Also, please let him know if you will be attending the luncheon.

I am looking forward to seeing you in November.

Sincerely,

Kenneth J. Lanfear

Kenneth J. Lanfear
President

cc: John J. Warwick, Editor, *Journal of the American Water Resources Association*
Janet L. Bowers, Chairperson, 2002 Awards/Nominations Committee

AMERICAN WATER RESOURCES ASSOCIATION
4 WEST FEDERAL STREET | P.O. BOX 1626 | MIDDLEBURG, VA 20118-1626
PHONE: 540.687.8390 | FAX: 540.687.8395 | WWW.AWRA.ORG

Center for Statistical Ecology and Environmental Statistics

PREDICTABILITY OF SURFACE WATER POLLUTION LOADING IN PENNSYLVANIA USING WATERSHED-BASED LANDSCAPE MEASUREMENTS

Glen D. Johnson[1,3], W. L. Myers[2,3], and Patil, G. P.[1,3]

[1]Center for Statistical Ecology and Environmental Statistics
Department of Statistics
[2]School of Forest Resources and Environmental Resources Research Institute
Environmental Resources Research Institute
[3] Intercollege Graduate Degree Program in Ecology

The Pennsylvania State University
University Park, PA 16802

NSF/EPA Project Officer: Penelope Firth and Barbara Levinson
EPA Project Officer: Barry D. Nussbaum

Prepared with partial support from the National Science Foundation Cooperative Agreement Number DEB-9524722 and the United States Environmental Protection Agency Cooperative Agreement Number CR-825506. The contents have not been subjected to Agency review and therefore do not necessarily reflect the views of the Agencies and no official endorsement should be inferred.

Journal of the American Water Resources Association
Volume 37, pages 821—835, 2001
Best Paper Award, 2002

Technical Report Number 99-0303
TECHNICAL REPORTS AND REPRINTS SERIES
August 1999

Department of Statistics
The Pennsylvania State University
University Park, PA 16802

G. P. Patil
Distinguished Professor and Director
Tel: (814)865-9442 Fax: (814)865-1278

Predictability of Surface Water Pollution Loading in Pennsylvania using Watershed-based Landscape Measurements

Glen D. Johnson[*,†,‡], Wayne L. Myers[§,‡],

and Ganapati P. Patil[†,‡]

[†]Center for Statistical Ecology and Environmental Statistics,

Department of Statistics,

[§]School of Forest Resources and

Environmental Resources Research Institute

[‡]Inter-college Graduate Degree Program in Ecology

Penn State University, University Park, PA 16802 USA

Prepared with partial support from the National Science Foundation Cooperative Agreement Number DEB-9524722 and the United States Environmental Protection Agency Cooperative Agreement Number CR-825506. The contents have not been subjected to Agency review and therefore do not necessarily reflect the views of the Agencies and no official endorsement should be inferred.

[*]Glen D. Johnson is currently a Research Scientist with the New York State Department of Health, Bureau of Environmental and Occupational Epidemiology, in the Geographic Research and Analysis Section, 547 River Street, Troy, New York 12180-2216

abstract

We formally evaluated the relationship between landscape characteristics and surface water quality in the state of Pennsylvania (USA) by regressing two different types of pollutant responses on landscape variables that were measured for whole watersheds. One response was the monthly exported mass of nitrogen estimated from field measurements, while the other response was a GIS-modeled pollution potential index. Regression models were built by the stepwise selection protocol, choosing an optimal set of landscape predictors. After factoring out the effect of physiography, the dominant predictors were the proportion of "annual herbaceous" land and "total herbaceous" land for the nitrogen loading and pollution potential index, respectively. The strength of these single predictors is encouraging because the marginal land cover proportions are the simplest landscape measurements to obtain once a land cover map is in hand; however, the optimal set of predictors also included several measurements of spatial pattern. Thus, for watersheds at this general hierarchical scale, gross landscape pattern may be an important influence on in-stream pollution loading. Overall, there is strong evidence that using landscape measurements alone, obtained solely from remotely sensed data, can explain most of the water quality variability ($R^2 \approx 0.75$) within these watersheds.

key terms: landscape patterns, nutrient loading, pollution potential, water quality, watersheds, multi-scale relationships

1 Introduction

Ecological heirarchy theory establishes a framework for explaining how large-scale characteristics of ecosystems can constrain smaller-scaled characteristics (Urban, O'Neill and Shugart, 1987; O'Neill, Johnson and King, 1989). An example of such an inter-scale environmental relationship is the influence of gross land use characteristics on local surface water quality. Indeed, for all the improvements in water quality associated with modern controls on point-source discharges, local water quality is still constrained by non point-source pollution. Since land use is generally reflected by land cover (vegetation type), then whole watersheds may be evaluated with respect to water quality risk by characterizing land cover proportions and patterns (O'Neill, et al., 1997). Watershed-wide landscape characteristics that are significantly correlated with local water quality may then serve as landscape-scale indicators of environmental condition, as also sought by other researchers (Aspinall and Pearson, 2000; Jones, et al, 1997).

A common theme that appears to arise from previous research in this area is that as a watershed gets larger, corresponding to higher order drainage basins, land cover proportions alone explain most of the water quality variability; whereas for smaller watersheds, especially those for first order headwater streams, the spatial pattern of land cover becomes more important (Graham, et al., 1991; Hunsaker and Levine, 1995; Roth, Allan and Erickson, 1996). This indicates that the feasibility of using watershed-wide marginal land cover proportions and/or spatial pattern measurements for predicting water quality depends on the hierarchical scale of a watershed.

This article presents an evaluation of the relationship between surface water pollution loading and landscape characterstics for watersheds in the state of Pennsylvania (USA) that are each about $1/100^{\text{th}}$ the size of the state. Using data from previous studies, linear models were developed for choosing an optimal set of landscape predictors that constituted both land cover proportions and pattern measurements.

2 Surface Water Pollution Assessment

2.1 Nitrogen Loading

A recent study (Nizeyimana, et al., 1997) was conducted to assess surface-water nutrient loading in Pennsylvania watersheds. The primary purpose was to quantify the various sources of non-point source (NPS) nutrient loading. Watersheds, as seen in the top of + 5, were delineated by choosing 85 Water Quality Network stations throughout Pennsylvania, then aggregating detailed sub-watershed boundaries that were previously digitized by the United States Geologic Survey. Each resulting NPS watershed then drains to one of the 85 network stations. As part of this study, total levels of both Nitrogen and phosphorous were obtained for each watershed by applying flow-weighted averaging techniques to monthly in-stream concentrations from the previous 5 years. The result is an estimate of the monthly exported mass in kilograms (kg).

[Figure 5 about here]

Meanwhile, Johnson, et al. (in press) obtained landscape measurements on a different set of watersheds that are based on the state water plan, as discussed in section 3 and delineated in the bottom of Figure 5. Since there was not a perfect matching between the two watershed delineations, some of the NPS watersheds were aggregated to equal the area of a state water plan watershed, as identified in the top of Figure 5. These watersheds are then added to those for which there is an exact or very close match with the state water plan-based watersheds and the final set are shaded in the bottom of Figure 5. The result is a sample of 30 watersheds across the state for which we have measurements of both landscape pattern and nutrient loading.

For the NPS watersheds that were aggregated, the nutrient loading was summed. All watershed-based estimates of total nitrogen and total phosphorous, in kilograms (kg), were divided by the total area, in hectares (ha), in order to adjust for the varying watershed sizes. Nitrogen was then plotted against phosphorous, as seen in Figure 2. Clearly, one only needs to pursue either nitrogen or phosphorous as an indicator of nutrient loading since they are so highly linearly correlated with each other. Therefore nitrogen was chosen because it is suspected to yield better quality data than phosphorous. This suspicion arises because nitrogen loading is always reported well above zero (minimum for these 30 watersheds = 27.12 kg/ha), whereas phosphorous loading is sometimes reported at less than 1 kg/ha, thus indicating that there were likely to be more measurements near or below analytical detection limits in the original water quality network samples.

[Figure 2 about here]

Since the objective of this study is to evaluate the effect of land use patterns on surface-water nutrient loading, it was considered to subtract the portion of total nitrogen loading that was estimated by Nizeyimana, et al. (1997) to be attributed to atmospheric deposition. However, of the two primary components of atmospheric nitrogen, ammonium (NH_4) was determined to come almost entirely from volatilization from manure and other fertilizers; while the other primary component, nitrogen oxides (NO_x) was determined to have about one third contributed by manure and other fertilizers and about two thirds from industrial/urban sources. Also, natural sources of atmospheric deposition of nitrogen was considered negligible (Nizeyimana, et al., 1997). Therefore, since much of the source of atmospheric nitrogen deposition can be attributed to local land use activity and natural "background" sources are negligible, total nitrogen loading was kept intact. A thematic presentation of total nitrogen loading is seen in Figure 5, along with the physiographic provinces of Pennsylvania, where the major provinces are labeled.

[Figure 5 about here]

2.2 Pollution Potential Index

Pennsylvania watersheds that are delineated by the state water plan watershed boundaries were evaluated in an earlier study (Hamlett, et al., 1992) through GIS modeling for ranking each watershed for its nonpoint source pollution potential. Various statewide data layers (coverages) were analyzed to produce four different indexes; a runoff index (RI), a chemical use index (CUI), a sediment production index (SPI) and an animal loading index (ALI). An overall pollution potential index (PPI) was then computed for each watershed by:

$$\mathrm{PPI_i} = W_1(\mathrm{RI_i}) + W_2(\mathrm{SPI_i}) + W_3(\mathrm{ALI_i}) + W_4(\mathrm{CUI_i}) \qquad (1)$$

for the i^{th} watershed, where W_1 to W_4 are weights assigned to each input index. The results represent per-acre average values. Petersen, et al. (1991) show results for an unweighted version of Equation 1 ($W_j = 0.25$ for $j = 1, \cdots, 4$) and a weighted version where the chemical use index is weighted downward to $W_4 = 0.10$ and the remaining input indexes were equally weighted at 0.30. Also, since the model depends heavily on land cover types, results were presented for both "agricultural land" and "all land". While the purpose of the initial study was to evaluate "agricultural" pollution potential, the purpose of the study being reported in this paper is to evaluate overall pollution potential. Therefore, we are fortunate that results were also presented by Petersen, et al. (1991) for "all lands".

Using the "equally weighted all lands" category, the resulting ranking of the watersheds are presented thematically in Figure 5. For graphical display and regression modeling, the ranks are presented in reverse of how they are reported by Petersen, *et al.* (1991) so that the increasing pollution potential is represented by increasing numerical value. The watersheds are stratified geographically in Figure 5 by aggregating physiographic sections, which are nested within physiographic provinces, in order to form more homogeneous areas with respect to PPI ranks.

[Figure 5 about here]

The original state water plan delineation, for which PPI values were obtained, consists of 104 watersheds; however, the delineation used for obtaining landscape measurements consists of 102 watersheds resulting from a more spatially accurate aggregation of smaller watersheds that were in turn originally digitized by the USGS. Two of the USGS-source watersheds each consist of two state water plan watersheds; therefore, out of the resulting 102 USGS-source watersheds, two of them did not have direct PPI assessments. For this reason, analysis was limited to 100 of the USGS-source watersheds for which both PPI values and landscape measurements were available. The two "missing" watersheds are indicated by diagonal hatching in Figure 5.

3 Selecting an Initial Set of Landscape Pattern Variables

In a separate study (Johnson, *et al.*, in press), landscape variables were measured for 102 of the state water plan-based watersheds through application of the FRAGSTATS software (McGarigal and Marks, 1995).

Land cover data, from which measurements were obtained, consisted of an 8-category raster map of Pennsylvania that was in turn derived from LANDSAT TM images with a pixel resolution of 30 meters. Details of how the raw satellite data was processed to derive the raster maps is available through metatdata located at the Pennsylvania Spatial Data Access web page (http://www.pasda.psu.edu), under the category of "Terrabyte images". The method and software is available as C-language programs for general use under the acronym PHASES (Myers, 1999).

The land cover categories are water, conifer forest, mixed forest, broadleaf forest, transitional, perennial herbaceous, annual herbaceous and terrestrial unvegetated. The category of transitional land derives from a heterogeneous mix of land cover types; perennial herbaceous is primarily grassland that occurs in small patches just about everywhere, but occurs in larger patches where pastureland is present; annual herbaceous is primarily cropland and is often adjacent to patches of perennial herbaceous land; meanwhile, terrestrial unvegetated is primarily urbanized land. The remaining category labels are self explanitory++. As listed at the end of Table 1, marginal (non-spatial) land cover measurements that were included for this study are a summation of all three forest types, then both herbaceous types and finally terrestrial unvegetated land cover.

[Table 1 about here]

A new multi-resolution characterization of spatial pattern, termed a conditional entropy profile (Johnson, Tempelman and Patil, 1995; Johnson and Patil, 1998; Johnson, et al., 1998 and 1999) was also obtained for all of the state water plan-based watersheds (Johnson, et al., in press). These profiles quantify landscape fragmentation by measuring entropy of the spatial distribution of land cover categories at a given raster map resolution in a way that is conditional on the categories of a coarser-resolution map. When computed for multiple resolutions, ranging from the "floor" that is provided by the original raster map to a resolution beyond which conditional entropy does not change much, a profile is traced out that reflects aspects of the underlying spatial pattern. Increasingly degraded resolutions are obtained by a resampling filter. An example profile and its parameterization is seen in Figure 5. Basically, A is the extent of information that is lost from degrading the map resolution, B is the rate of information loss and C is the asymptotic conditional entropy that is highly correlated with the entropy of the marginal (non-spatial) land cover distribution.

[Figure 5 about here]

A set of variables was sought that show little to no correlation among themselves in order to avoid multicollinearity in regression modeling. Therefore, an approximately orthogonal subset of spatial pattern variables was obtained by applying principal components analysis to the full set of pattern variables in Table 1 along with non-linear regression estimates of the conditional entropy profile parameters A,B and C. The marginal land cover proportions were excluded from this data reduction exercise because it was desired to include all of the land cover proportions in the set of potential predictors. Since this set of variables consists of differing measurement units, eigen analysis was performed on the correlation matrix. Results for the 30 watersheds that shared both landscape and nitrogen loading measurements are presented here. When re-applied to all of the 102 watersheds for which there are landscape measurements, the results were essentially the same and are therefore not reproduced here.

As seen in Figure 6, the first four components explained over 90% of the variability in the original multivariate data set. Correlations between the original variables and the principal components, which are simply the eigenvector elements (loadings) multiplied by the square root of the corresponding eigenvalue (Stiteler, 1979), are reported in Table 2.

[Figure 6 about here]

[Table 2 about here]

The first component is very highly correlated with those variables that are in turn highly correlated with the marginal land cover distribution. This component reveals the contrast between watersheds that tend towards being more fragmented and more evenly distributed with smaller patches (positive loadings) and those with a high degree of patch coherence (negative loadings). Although many of the original variables could be chosen for representing the first component, contagion (CONTAG) was chosen because it is the most highly correlated and is a very familiar measurement in landscape ecology.

The second component is mostly correlated with the conditional entropy profile parameter estimates A and B (note that C is highly correlated with component 1, as expected). This component contrasts high values of A, and secondarily the landscape shape index (LSI), as reflected by positive loadings, with high values of B, as reflected by negative loadings.

The third component is most highly correlated with the fractal dimension characterization of patch shape (DLFD) and secondarily with the patch size coefficient of variation (PSCV). Meanwhile, the fourth component is dominated by the landscape shape index.

In view of the results of principal components analysis, the spatial pattern variables that were included in the set of potential regressors were patch size coefficient of variation (PSCV), landscape shape index (LSI), fractal dimension (DLFD), contagion (CONTAG) and the conditional entropy profile values A and B. Finally, the proportions of annual herbaceous land (ANN.HERB), total herbaceous land (TOT.HERB), which is the sum of annual and perennial herbaceous land, and total forest land (TOT.FOREST), which is the sum of broadleaf, conifer and mixed forest lands, were added to the set of potential regressors.

Relationships among the final set of potential landscape predictors for the sample of 30 watersheds containing both landscape and nitrogen loading measurements are seen in Figure 7, where total nitrogen is also included as a log transform (logN) for reasons discussed later. One expects the proportion of annual herbaceous land to be a very strong, if not dominant, predictor of total nitrogen loading since it consists mostly of cropland. Agriculture was determined to be a main source of nitrogen loading in the initial study (Nizeyimana, et al., 1997).

[Figure 7 about here]

Meanwhile, relationships among the variables for all of the 102 watersheds are presented in Figure 8 along with the inverse of the PPI rank (PPI.INV).

[Figure 8 about here]

The different land cover proportions plotted in Figures 7 and 8 are highly inter-correlated, as expected, and some redundancy exists between PSCV and CONTAG as well as between the values of A and B; however, it is desired to include all of these variables in the initial set of landscape measurements in order to see which may be chosen over others as part of a stepwise model building protocol, as discussed in section 4.

4 Linear Models for Relating Water Pollution Loading to Landscape Variables

Stepwise regression was applied separately for each response variable–total nitrogen and pollution potential index–in order to build an optimal linear model from the potential set of regressors in Figures 7 and 8. The criterion for choosing the best set of predictors was a modification of Mallow's Cp statistic (Mallows, 1973), known as the Akaike Information Criterion (AIC) (Akaike, 1974). The AIC is related to the Cp statistic by the relation

$$\mathrm{AIC} = \hat{\sigma}^2(\mathrm{Cp} + n),$$

for n observations and $\hat{\sigma}^2$ equals the mean squared error of the initial model before adding or deleting a term to yield the "new" p-parameter model (MathSoft, Inc., 1997, p. 132). The result is

$$\mathrm{AIC} = RSS(p) + MSE * 2 * p \quad , \tag{2}$$

where $RSS(p)$ is the residual sum of squares from the new model defined by p terms (k predictors plus the intercept) and MSE is the mean squared error from the original model prior to deleting or adding a term.

The automated stepwise selection protocol works by choosing the set of predictors that minimizes the AIC statistic. Critical F values for deciding whether or not to include or remove predictor variables was set at 2, subsequently erring in favor of retaining large sets of predictor variables.

Models were checked by the usual diagnostic graphics. In addition, partial residual plots were obtained for each regressor in a model. Following Montgomery and Peck (1982), the i^{th} partial residual for the regressor x_j is

$$\begin{aligned} e^*_{ij} &= y_i - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_{j-1} x_{i,j-1} - \hat{\beta}_{j+1} x_{i,j+1} - \cdots - \hat{\beta}_k x_{ik} \\ &= e_i + \hat{\beta}_j x_{ij} \qquad \text{for } \; i = 1, \cdots, n. \end{aligned} \tag{3}$$

These partial residual plots display the relationship between y and the regressor x_j after the effect of the other regressors $x_i (i \neq j)$ have been removed, therefore more clearly showing the influence of x_j, given the other regressors. Along with providing a check for outliers and inequality of variance, these plots also indicate more precisely how to transform the data to achieve linearity than do the usual residual plots.

4.1 Predicting In-Stream Nitrogen Loading

Initial analysis was performed using total nitrogen (kg/ha) as the response variable; however, the resulting model was excessively influenced by two watersheds from the Piedmont physiographic province (see Figure 5). A natural log transform substantially reduced the domineering influence of these two watersheds and yielded other diagnostics that were much better; therefore, all analyses proceeded with the log of total nitrogen (logN) as the response variable.

The graphical relationship of logN with the potential predictor variables is seen in Figure 7. Although this shows a fairly strong linear relationship between logN and the marginal land cover measurements, that in turn all appear highly correlated among themselves, some of the other potential predictors may also explain a significant portion of the variability in observed logN. Actually, scatter diagrams can be misleading in the case of multiple regression, as pointed out by Montgomery and Peck (1982, p. 122), who cite Daniel and Wood (1980).

Preliminary analysis showed that when all 30 of the NPS watersheds were included, the only variable retained by the stepwise selection procedure was the proportion of annual herbaceous land (ANN.HERB); however, when separate analyses were performed within each major physiographic province, very different results were obtained. For the 12 watersheds of the Appalachian Plateaus, all but the landscape shape index were retained. Since the Ridge and Valley had only 7 NPS watersheds, they were combined with the Piedmont, which reveals similar forest fragmentation patterns. For the 18 watersheds of the combined Piedmont / Ridge and Valley Province group, annual herbaceous land was retained along with the fractal dimension (DLFD) and both the A and B values of the conditional entropy profiles.

Upon seeing large differences in the resulting models, given the physiographic region, and desiring to maximize the residual degrees of freedom associated with any final model, the analysis was continued by combining all 30 watersheds from across the state and including an indicator (0,1) variable (sometimes called a dummy variable) for designating membership in a physiographic region. The indicator variable, which was forced to be retained by the stepwise protocol, was coded with 1 (one) if the corresponding watershed was from the Piedmont/Ridge and Valley group, and a 0 (zero) otherwise. The resulting parameter estimate revealed the increase (or decrease) in total nitrogen loading as one moves from the Appalachian Plateaus to the Piedmont/Ridge and Valley group. A further advantage of factoring out pyhsiographic regions is to reduce deleterious effects of possible spatial autocorrelation.

Coefficient estimates for the model that minimized the AIC statistic (AIC = 2.84) are reported in Table 3, where the dummy variable indicating the effect of province group is labeled as PIED.RV.

[Table 3 about here]

Diagnostic plots for the model defined in Table 3 revealed a strong linear relationship between the fitted and observed values, along with randomly scattered residuals. A Q/Q plot revealed somewhat heavy tails in the distribution of residuals; however, none of these observations were excessively influential according to Cook's Distance. Generally, a Cook's Distance of 1 or greater is considered to reveal an overly influential observation (Montgomery and Peck, 1982; Neter, Wasserman and Kutner, 1985) which is far greater than the worst case. These diagnostics therefore revealed a very acceptable model.

The partial residual plots for each quantitative predictor in Table 3 appear in Figure 9 where the lines of fit have slopes equal to the parameter estimates in Table 3. The plots in Figure 9 indicate a linear trend for each predictor, especially for annual herbaceous land (ANN.HERB), and no data transformations appear to be necessary.

[Figure 9 about here]

All possible interactions between the quantitative variables and the indicator variable were investigated, but none of these interactions turned out to be at all significant. When interactions were evaluated among the quantitative variables, the two-way interaction between LSI and ANN.HERB was significant ($p = 0.025$). However, when the model parameters were re-computed, including LSI*ANN.HERB as the only interaction term, the estimate of the ANN.HERB coefficient became negative, which is nonsense. Therefore, the initial additive model in Table 3 was retained.

spatial autocorrelation

The presence of spatial autocorrelation was evaluated by plotting residuals from the model defined in Table 3 as a function of geographic distance of the center of each watershed from the center of the watershed yielding the maximum residual. Selecting the initial watershed (distance = 0) is rather arbitrary, but it was felt that the most likely trend would be a general decrease in nitrogen loading as one moves away from a "hot-spot" watershed; therefore starting with the watershed yielding the maximum residual may help distinguish such a downward trend. Finally, since the indicator variable Pied.RV already serves to factor out a major spatial component, distance measurements were made within each of the two physiographic province groups.

Figure 10 plots the residuals as a function of distance. Keep in mind that an initial downward trend will always occur between the initial watershed and the next closest one since the initial one was chosen from yielding the largest residual; therefore, focus should be on all but the initial watershed. The Piedmont/Ridge and Valley did not visually reveal any spatial dependence, which was quite encouraging; however, the Appalachian Plateaus did reveal a downward trend that was followed by an upturn. This quadratic type response is due to a downward trend as one moves from watersheds near Pittsburgh on northward through mixed agricultural areas, then eastward to mostly forested areas, then further eastward to the Pocono region along the Delaware River.

[Figure 10 about here]

As an attempt to overcome the autocorrelated residuals in the Appalachian

Plateaus, watersheds in this physiographic province were further separated into two groups according to the finer scaled physiographic sections. After forcing two indicator variables to be retained for representing three spatial groups, the stepwise selection protocol yielded a similar model to that in Table 3 with the exception that fractal dimension (DLFD) was replaced by contagion (CONTAG).

Diagnostic plots, however, indicated that model quality had somewhat decreased. Furthermore, small sample sizes within each of the newly defined groups of physiographic sections within the Appalachian Plateaus physiographic province made it difficult to truly discern any residual autocorrelation. Therefore, the model in Table 3 was chosen. One should consider, however, that the mean squared error may slightly underestimate the true variance due to some positive spatial autocorrelation.

4.2 Predicting a Pollution Potential Index

Unlike with nitrogen loading, the pollution potential data can be treated as observations on a population of watersheds (100 out of 102). Since the computed linear coefficients are actually parameter values, standard errors are not relevant and thus are not reported. However, it is still sensible to choose an optimal set of predictors by minimizing the AIC statistic. Furthermore, t-scores and p values are still reported in order to see the significance of each term, relative to the other terms.

For the purpose of regression modeling, the five geographic strata that appear in Figure 5 are represented by four indicator variables that are explained in Table 4. These indicators were forced to be retained by the model selection protocol in order to factor out physiographic effects and minimize possible spatial autocorrelation. The resulting parameter estimates reveal the increase (or decrease) in average PPI rank as one moves from the "Pittsburgh Plateaus/Allegheny Mountains” group to the group being represented by the respective indicator variable. The model that minimized the AIC statistic is presented in Table 4.

[Table 4 about here]

Diagnostic plots for the model defined in Table 4 revealed a very strong linear relationship between the fitted and observed values, along with thoroughly randomly scattered residuals. A Q/Q plot revealed that the residuals are closely approximated by the normal distribution. Further, none of these observations are excessively influential according to Cook’s Distance. Consequently, these diagnostics reveal a very acceptable model.

The partial residual plots for each quantitative predictor in Table 4 appear in Figure 11. The lines of fit in Figure 11 have slopes equal to the parameter estimates in Table 4. The plots in Figure 11 indicate a linear trend for each predictor, and no data transformations appear to be necessary.

[Figure 11 about here]

5 Interpretation

The chosen model for relating total nitrogen loading (kg/ha) to landscape characteristics within Pennsylvania watersheds that are delineated based on the state water plan is as follows:

$$\begin{aligned} \ln(\mathrm{N}) &= 12.78 + 0.42(\mathrm{Pied.RV}) + 3.24(\mathrm{ANN.HERB}) \\ &\quad + 0.0034(\mathrm{LSI}) - 5.89(\mathrm{DLFD}) - 0.41(\mathrm{A}), \end{aligned} \tag{4}$$

where the associated statistics for the parameter estimates based on a sample of 30 watersheds, and an explanation of the variable labels are found in Table 3. The associated variance σ^2 is estimated by MSE=0.075, although this might be a slight underestimate due to some spatial autocorrelation in the Appalachian Plateaus.

As expected, the dominant regressor is the proportion of annual herbaceous land which, in turn, is mainly cropland. Further, given the proportion of annual herbaceous land and physiographic membership, landscape pattern strengthened the explanation of nutrient loading variability among these Pennsylvania watersheds, as measured through total nitrogen loading. The landscape shape index (LSI), the fractal dimension estimate (DLFD) and the estimate of conditional entropy profile "depth" (A) were all retained by the stepwise selection procedure which aims to minimize the residual sum of squares and corresponding AIC statistic out of all possible regressions.

The slight, but significant, positive relation to the landscape shape index indicates that nitrogen loading may be expected to increase as the landscape becomes more fragmented, resulting in more edges.

A negative relation to the value "A" is not readily interpretable; however, it is noteworthy that this predictor and LSI were both always retained by the stepwise selection procedure whether the physiography indicator variables were designed to differentiate among the 3 major provinces (results not shown here), the 2 province groups (Appalachian Plateaus vs. Piedmont/Ridge and Valley) or the 3 groups that consisted of the Piedmont/Ridge and Valley and 2 sub-areas of the Appalachian Plateaus.

The chosen model for relating the pollution potential index (PPI) rank to landscape characteristics within Pennsylvania watersheds that are delineated based on the state water plan is as follows:

$$\begin{aligned}\text{PPI rank} &= 445.2 - 6.3(\text{APP. MOUNTAIN}) \\ &\quad +15.72(\text{PIED. and GR. VALLEY}) \\ &\quad -2.35(\text{LOW and POCONO}) - 12.43(\text{HIGH PLATEAUS}) \\ &\quad +118.9(\text{TOT.AG}) - 330.2(\text{DLFD}) - 1.0(\text{CONTAG}) \\ &\quad +27.8(\text{A}) + 110.4(\text{B}), \qquad (5)\end{aligned}$$

where an explanation of the variable labels is found in Figure 8 and Table 4.

The dominant regressor is the proportion of total herbaceous land; however, results show that given the proportion of total herbaceous land and physiographic membership, that landscape pattern still strengthens the explanation of surface water pollution potential variability among these Pennsylvania watersheds.

The negative relation to fractal dimension is consistent with the nitrogen loading results. A negative relation makes sense because when landscape patches are left to natural forces, they tend towards having more irregular outlines, which is reflected by an increasing fractal dimension (or perimeter/area scaling exponent) (Johnson, Tempelman and Patil, 1995), while patches that are created and maintained by humans tend to have straight edges, especially with cropland that is in turn largely responsible for nutrient loading. As the average landscape patch tends towards having a straighter edge, this is reflected by a fractal dimension estimate that tends towards a value of 1, the dimension of a Euclidean line. A negative relation to contagion is likely due to the highest levels of contagion being associated with mostly forested watersheds. Although both conditional entropy profile variables A and B are retained by the stepwise protocol, a mechanistic explanation of their relation to PPI is not necessarily clear.

As an exploratory exercise, nine watersheds were chosen to include the top three, middle three and lowest three nitrogen loading values, and this was repeated for the PPI values. Their corresponding conditional entropy profiles appear in Figures 12 and 13. For both nitrogen loading and the PPI, the three least polluted watersheds are clearly separate from the others which, in turn, are essentially grouped together. These three watersheds with the lowest pollution potential are mostly forested watersheds from the High Plateaus or Poconos and consistently reveal lower profiles that are "intrinsically less fragmented" than the other six profiles. Although these profiles do not reveal apparently large differences in A and B values, the model for predicting nitrogen loading benefitted from including A and the ability to predict pollution potential was strengthened from including both A and B.

[Figure 12 about here]

[Figure 13 about here]

In summary, the best landscape-level predictor of water pollution for these Pennsylvania watersheds is the marginal land cover proportions. A majority of nitrogen loading variability was explained by the proportion of annual herbaceous land, which is mostly row crops. Meanwhile, variability of the pollution potential index was largely explained by total herbaceous land, which includes annual and perennial herbaceous land. This finding agrees with results by Roth, et al.. (1996), who found that stream biotic integrity was significantly correlated with the proportion of agricultural land throughout a whole watershed. These authors further concluded that stream conditions are primarily determined by regional land use, overwhelming the ability of local riparian vegetation to support high quality habitat. Also, Hunsaker and Levine (1995) determined that nitrogen, phosphorous and conductivity were all primarily dictated by land use proportions and they further cite other studies that lead to similar findings. This is all quite encouraging because once a reliable land cover map is in place, the marginal land cover proportions are readily available; therefore, without any further information, one can make a fairly strong prediction of surface-water quality within a watershed.

We, however, found that additional measurements of spatial pattern for these watershed-delineated landscapes in Pennsylvania can significantly strengthen the predictability of pollution loading within the watershed. Furthermore, some aspects of the multi-resolution conditional entropy profiles were consistently retained by an objective variable selection protocol.

Literature Cited

Akaike, H. 1974. A new look at statistical model identification. IEEE Transactions on Automatic Control AU-19, pp. 716-722.

Aspinall, R. and Pearson, D. 2000. Integrated geographical assessment of environmental condition in water catchments: Linking landscape ecology, environmental modelling and GIS. J. Environmental Management, 59:299-319.

Daniel, C. and Wood, F.S. 1980. Fitting Equations to Data, 2nd Ed. Wiley, New York

Graham, R.L., Hunsaker, C.T., O'Neill, R.V. and Jackson, B.L. (1991). Ecological risk assessment at the regional scale. Ecological Applications, 1(2): 196-206.

Hamlett, J.M., Miller, D.A., Day, R.L., Peterson, G.W., Baumer, G.M. and Russo, J. 1992. Statewide GIS-based ranking of watersheds for agricultural pollution prevention. J. Soil and Water Conservation, 47(5):399-404.

Hunsaker, C.T. and Levine, D.A. 1995. Hierarchical approaches to the study of water quality in rivers. BioScience, 45(3):193-203.

Jones, K.B, Riitters, K.H., Wickham, J.D., Tankersley, R.D. Jr., O'Neill, R.V., Chaloud, D.J., Smith, E.R. and Neale, A.C. 1997. An Ecological Assessment of the United States Mid-Atlantic Region. USEPA, ORD, EPA/600/R-97/130.

Johnson, G.D., Myers, W.L, Patil, G.P. and Taillie, C. 1998. Quantitative characterization of hierarchically scaled landscape patterns. 1998 American Statistical Association Proceedings of the Section on Statistics and the Environment, pp. 63-69.

Johnson, G.D., Myers, W.L, Patil, G.P. and Taillie, C. 1999. Multiresolution fragmentation profiles for assessing hierarchically structured landscape patterns. Ecological Modeling, 116:293-301.

Johnson, G.D., Myers, W.L, Patil, G.P. and Taillie, C. in press. Categorizing and Monitoring Watershed-Delineated Landscapes in Pennsylvania using Conditional Entropy Profiles. Landscape Ecology.

Johnson, G.D. and Patil, G.P. 1998. Quantitative multiresolution characterization of landscape patterns for assessing the status of ecosystem health in watershed management areas. Ecosystem Health, 4(3):177-187.

Johnson, G.D., Tempelman, A.K. and Patil, G.P. (1995). Fractal based methods in ecology: a review for analysis at multiple spatial scales. Coenosis, 10(2-3), 123-131.

Mallows, C.L. 1973. Some comments on Cp. Technometrics, 15:661-675.

MathSoft, Inc.. 1997. Splus 4 Guide to Statistics. Data Analysis Products Division, MathSoft, Seattle. 877 pp.

McGarigal, K. and Marks, B. 1995. FRAGSTATS: spatial pattern analysis program for quantifying landscape structure. Gen. Tech. Rep. PNW-GTR-351. Portland, OR: U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station. 122 pp.

Montgomery , D.C. and Peck, E.A. 1982. Introduction to Linear Regression Analysis. John Wiley and Sons, New York, 504 pp.

Myers, W. 1999. Remote Sensing and Quantitative Geogrids in PHASES [Pixel Hyperclusters As Segmented Environmental Signals], Release 3.4. Technical Report ER9710, Environmental Resources Research Institute, The Pennsylvania State University, University Park, PA. 57 pp. and 2 diskettes.

Neter, J., Wasserman, W. and Kutner, M.H. 1985. Applied Linear Statistical Models, second ed. Richard D. Irwin, Homewood, Ill., 1127 pp.

Nizeyimana,E, et al.. 1997. Quantification of NPS Pollution Loads Within Pennsylvania Watersheds. ER9708. Environmental Resources Research Institute, University Park PA.

O'Neill, R.V., Hunsaker, C.T., Jones, K.B., Riitters, K.H.,Wickham, J.D., Schwartz, P.M.,Goodman, I.A., Jackson, B.L. and Baillargeon, W.S. 1997. Monitoring environmental quality at the landscape scale; using landscape indicators to assess biotic diversity, watershed integrity and landscape stability. Bioscience, 47(8):513-519.

O'Neill, R.V., Johnson, A.R. and King, A.W. (1989). A hierarchical framework for the analysis of scale. Landscape Ecology, 3(3/4), 193-205.

Petersen, G.W., Hamlett, J.M., Baumer, G.M, Miller, D.A., Day, R.L. and Russo, J.M. 1991. Evaluation of Agricultural Nonpoint Pollution Potential in Pennsylvania Using a Geographical Information System. ER9105. Environmental Resources Research Institute, University Park PA.

Roth, N.E., Allan, J.D. and Erickson, D.L. 1996. Landscape influences on stream biotic integrity assessed at multiple spatial scales. Landscape Ecology, 11(3):141-156.

Stiteler, W.M. 1979. Multivariate statistics with applications in ecology. in Orloci, L., Rao, C.R. and Stiteler, W.M (eds.). Multivariate Methods in Ecological Work. International Co-operative Publishing House, Fairland, Maryland. pp. 279-300.

Urban, D.L., O'Neill, R.V. and Shugart, H.H. Jr. 1987. Landscape ecology; a hierarchical perspective can help scientists understand spatial patterns. Bioscience, 37(2):119-127.

Table 1: Landscape variables measured for Pennsylvania watersheds.

Variable Description	Code
Patch Density	PD
Mean Patch Size	MPS
Patch Size Coefficient of Variation	PSCV
Edge Density	ED
Landscape Shape Index	LSI
Area-Weighted Mean Shape Index	AWMSI
Double-Log Fractal Dimension	DLFD
Area-Weighted Mean Patch Fractal Dimension	AWMPFD
Shannon Evenness Index	SHEI
Interspersion and Juxtaposition Index	IJI
Contagion†	CONTAG
Total Forest Cover	TOT.FOREST
Total Herbaceous Cover	TOT.HERB
Terrestrial Unvegetated	TU

note that diagonal pixels were included when determining patches

† pixel order preserved when measuring contagion

Table 2: Correlations between the original variables and the first five principal components.

variable	comp 1	comp 2	comp 3	comp 4	comp 5
PD	0.94	-0.18	0.20	0.07	-0.09
MPS	-0.93	0.22	-0.14	-0.03	0.14
PSCV	-0.78	0.02	0.53	0.16	-0.24
ED	0.94	-0.24	0.13	0.03	0.12
LSI	0.23	0.69	0.07	0.62	0.26
AWMSI	-0.84	-0.09	0.46	0.18	-0.12
DLFD	0.6	-0.04	0.63	-0.24	0.39
AWMPFD	-0.93	-0.01	0.31	-0.1	-0.04
SHEI	0.96	0.19	0.10	-0.07	-0.10
IJI	0.87	0.15	0.15	0.07	-0.38
CONTAG	-0.99	-0.02	-0.10	0.01	0.05
A	0.00	0.93	-0.20	-0.15	-0.11
B	0.26	-0.84	-0.3	0.3	-0.01
C	0.94	0.29	0.01	-0.02	-0.03

Table 3: Coefficients and corresponding statistics from regressing the log of total Nitrogen/ha against quantitative landscape variables and an indicator variable for specifying membership in a physiographic province group.

Regressor*	Value	Std. Error	t value	$\Pr(>\lvert t\rvert)$
Intercept	12.7805	4.9614	2.5760	0.0166
Pied.RV	0.4178	0.1950	2.1424	0.0425
LSI	0.0034	0.0015	2.3090	0.0299
DLFD	-5.8933	3.1854	-1.8501	0.0766
ANN.HERB	3.2441	0.8123	3.9936	0.0005
A	-0.4102	0.2271	-1.8058	0.0835

*Pied.RV reflects change due to membership in the Piedmont/Ridge and Valley group of physiographic provinces, relative to the Appalachian Plateau.

LSI = Landscape Shape Index, DLFD = Fractal Dimension estimate, ANN.HERB = proportion of Annual Herbaceous land, and A = estimate of conditional entropy profile depth.

mean squared error (24 d.f.) = 0.075; multiple R^2 = 0.74

Table 4: Linear coefficients from regressing the PPI rank against quantitative landscape variables and physiographic indicator variables. Mean squared error = 245.55 (90 d.f.) and multiple $R^2 = 0.76$.

Regressor*	Coefficient	t value	$\Pr(>\lvert t \rvert)$
Intercept	445.2282	1.79	0.0762
APP. MOUNTAIN	-6.2750	-1.09	0.2788
PIED. and GR. VALLEY	15.7200	2.44	0.0167
LOW and POCONO	-2.3516	-0.35	0.7252
HIGH PLATEAUS	-12.4292	-1.88	0.0638
DLFD	-330.2211	-2.46	0.0159
CONTAG	-0.9732	-2.19	0.0311
TOT.HERB	118.9058	5.01	0.0000
A	27.8007	2.26	0.0261
B	110.3800	2.61	0.0107

*Labels for the quantitative variables are explained in Figure 8

APP. MOUNTIAN = Appalachian Mountain Section

PIED. and GR.VALLEY = the Piedmont Plateau and Great Valley Section

LOW and POCONO = Glaciated Low and Pocono Plateau Sections

HIGH PLATEAUS = High and Mountainous High Plateau Sections

Figure 1: Watersheds from the NPS study (above), for which nutrient loadings are available, and watersheds based on the state water plan (below), for which landscape pattern measurements are available. Those NPS watersheds that can be aggregated to equal a state water plan watershed are indicated by grey above, and the final set of watersheds that have both landscape measurements and nutrient loading are in grey below.

Figure 2: Total Nitrogen vs Total phosphorous (kg/ha).

Figure 3: Thematic presentation of nitrogen loading in kilograms per hectare for 30 watersheds.

Figure 4: Thematic presentation of the pollution potential ranking for each of the state water plan-based watersheds. Physiographic stratification delineates more homogeneous geographic areas.

Figure 5: Anatomy of a conditional entropy profile.

Figure 6: Variance contributed by the first ten principal components; cumulative variance is labeled above each bar.

Figure 7: Pairwise scatterplots of the final set of potential predictor variables (regressors) along with the natural logarithm of total nitrogen per hectare (logN) for 30 watersheds. The spatial pattern variables are as follows: PSCV = Patch Size Coefficient of Variation, LSI = Landscape Shape Index, DLFD = Double Log Fractal Dimension, CONTAG = contagion and conditional entropy profile parameter estimates (A,B). Marginal land cover proportions are ANN.HERB = annual herbaceous, TOT.FOREST = total forest and TOT.HERB = total herbaceous.

Figure 8: Pairwise scatterplots of the set of potential predictor variables (regressors) along with the inverse of the pollution potential index (PPI.INV) for 102 watersheds. The landscape variables are explained in Figure 7.

Figure 9: Partial residual plots for the predictors listed in Table 3. Slopes of the fitted lines equal the parameter estimates in Table 3.

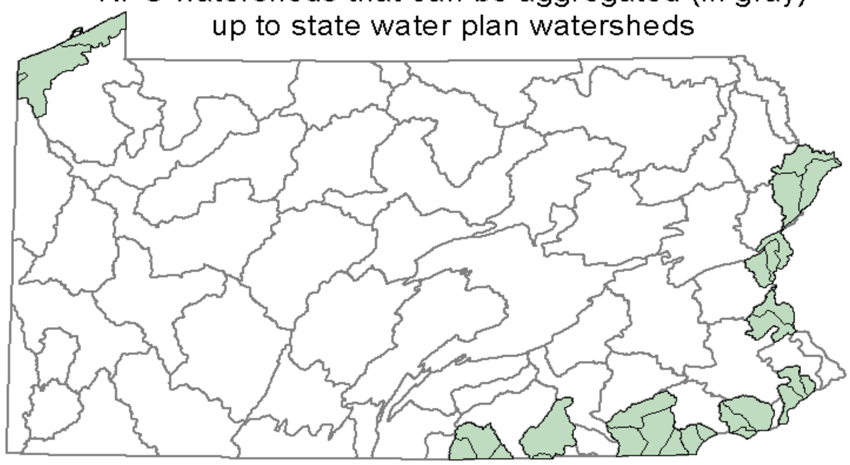
Figure 10: Residuals plotted as a function of geographic distance of the corresponding watershed from the watershed that yielded the maximum residual, given either the Appalachian Plateau or the Piedmont/Ridge and Valley physiographic province group.

Figure 11: Partial residual plots for the quantitative predictors listed in Table 4. Slopes of the fitted lines equal the linear coefficients in Table 4.

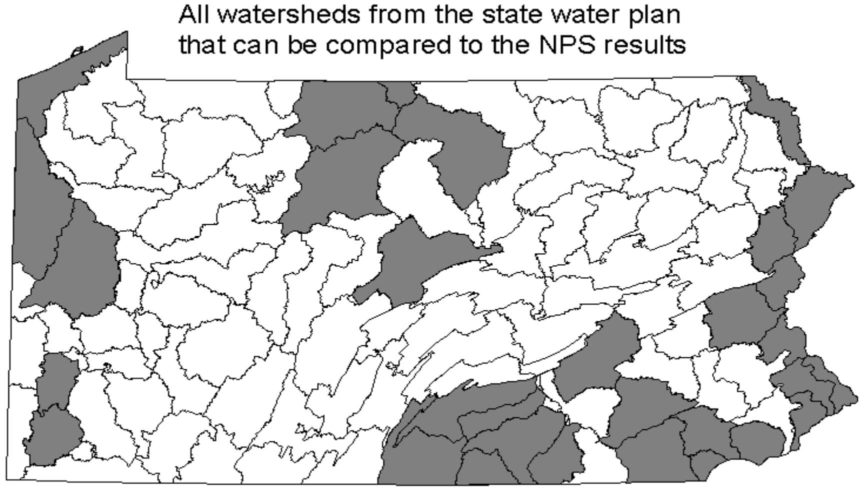
Figure 12: Conditional entropy profiles for watersheds containing the top 3, middle 3 and bottom 3 nitrogen loadings.

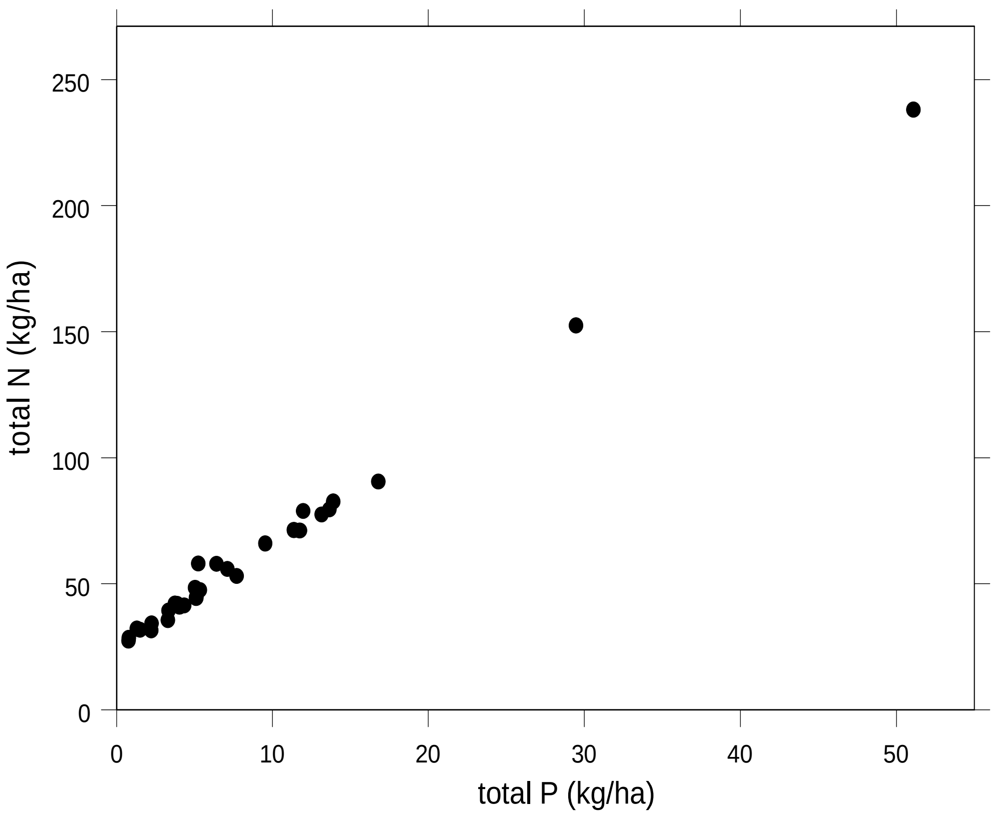
Figure 13: Conditional entropy profiles for watersheds containing the top 3, middle 3 and bottom 3 PPI values.

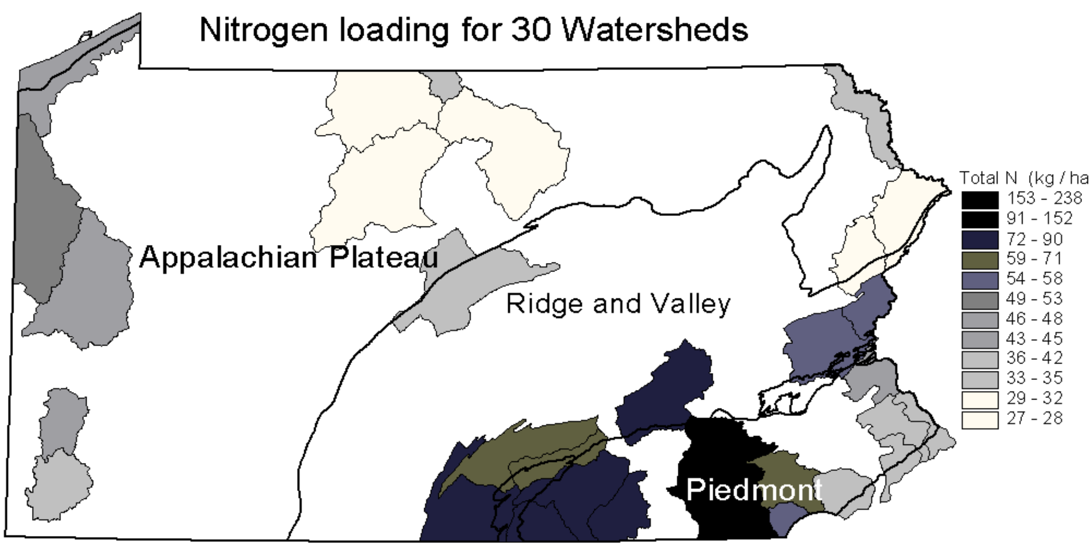
NPS watersheds that can be aggregated (in gray)
up to state water plan watersheds



All watersheds from the state water plan
that can be compared to the NPS results

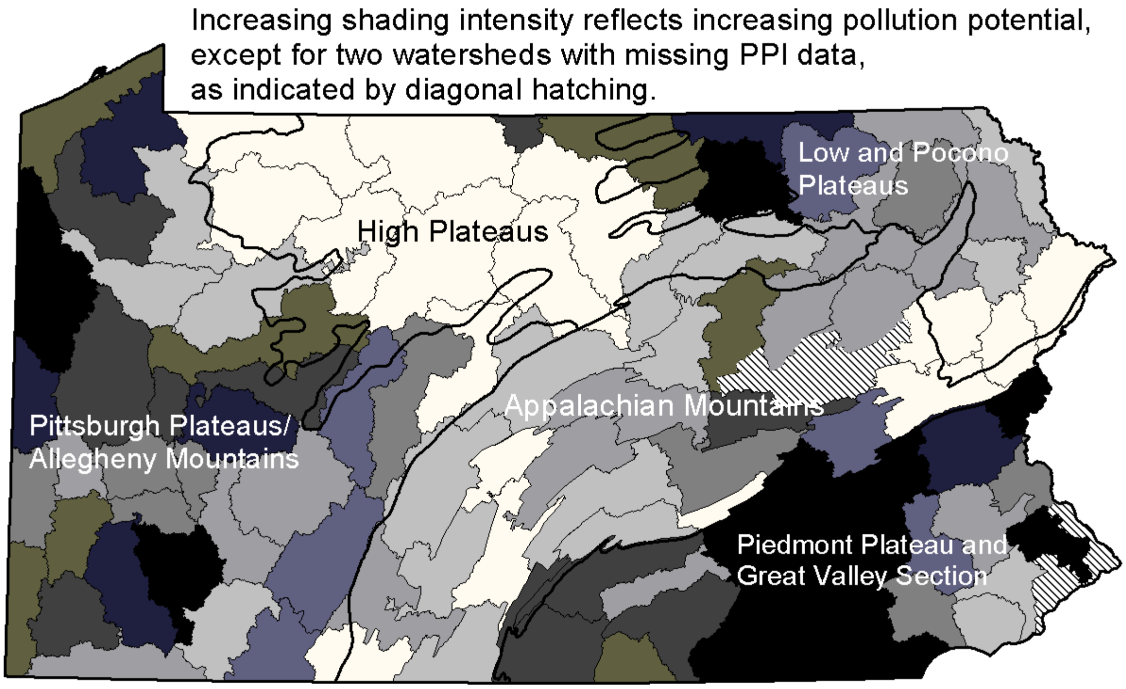



total N (kg/ha)
250
200
150
100
50
0
0
10
20
30
40
50
total P (kg/ha)

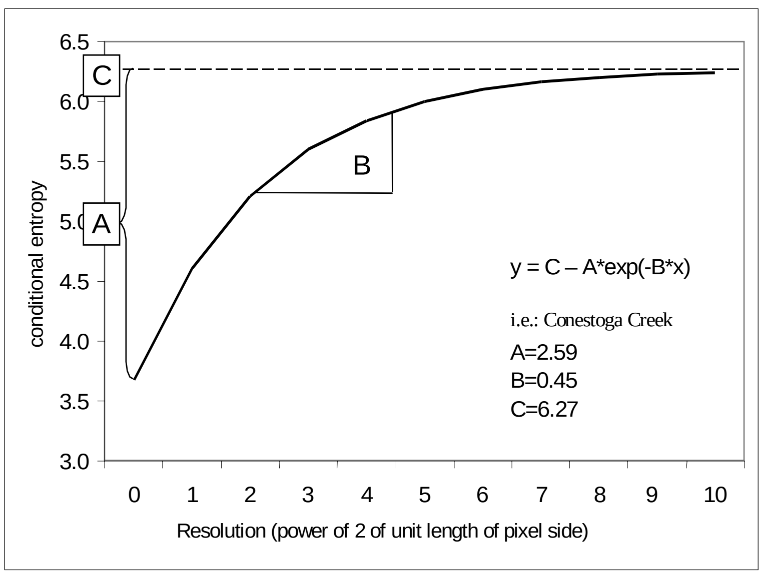


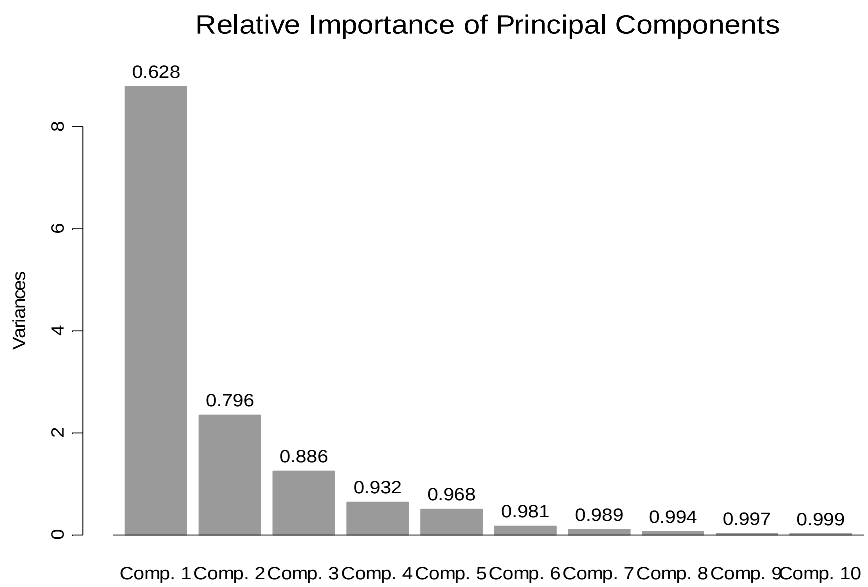
All state water plan-based watershed boundaries shown;
Physiographic Provinces delineated by heavy lines

Pollution Potential Index

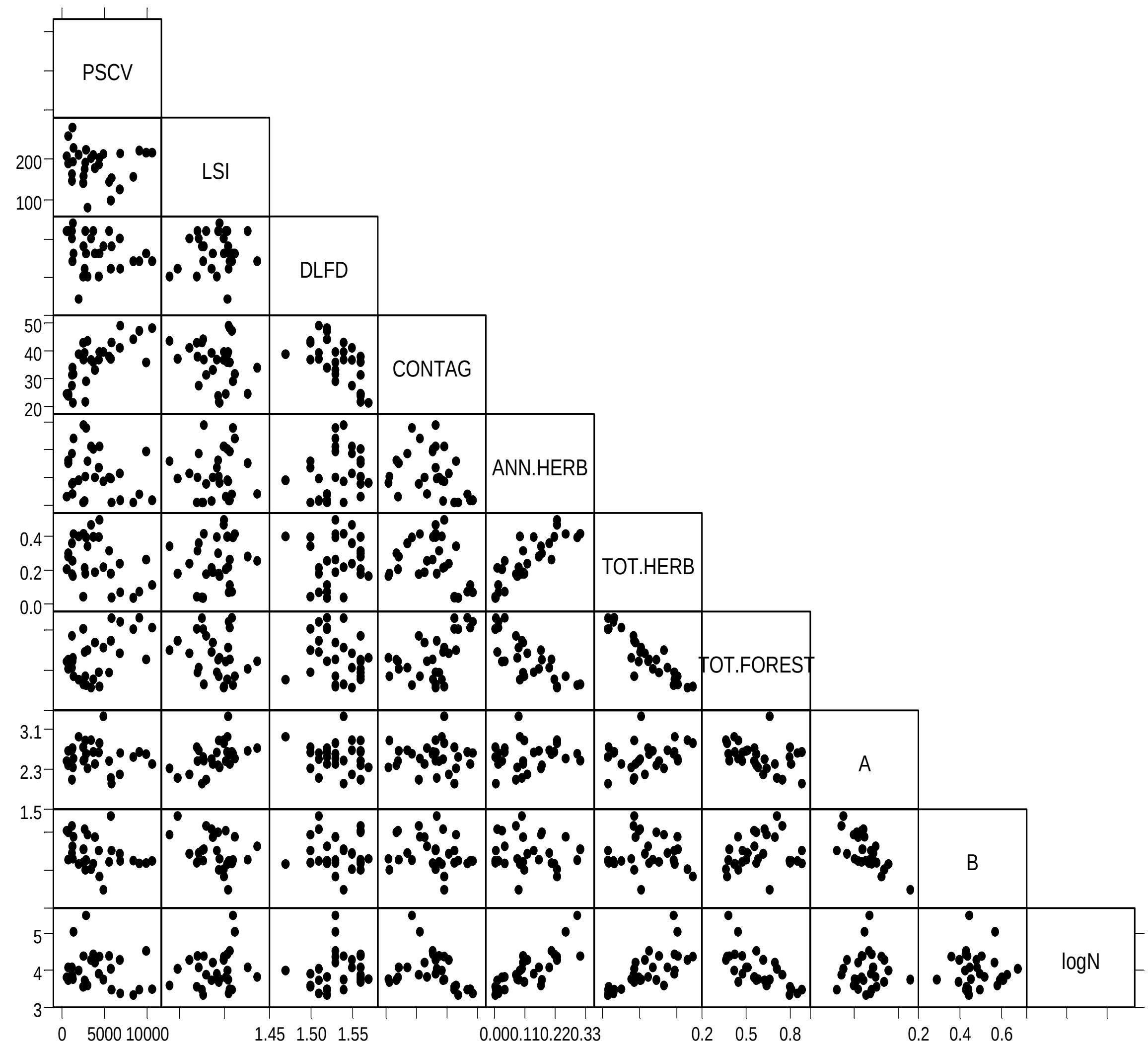


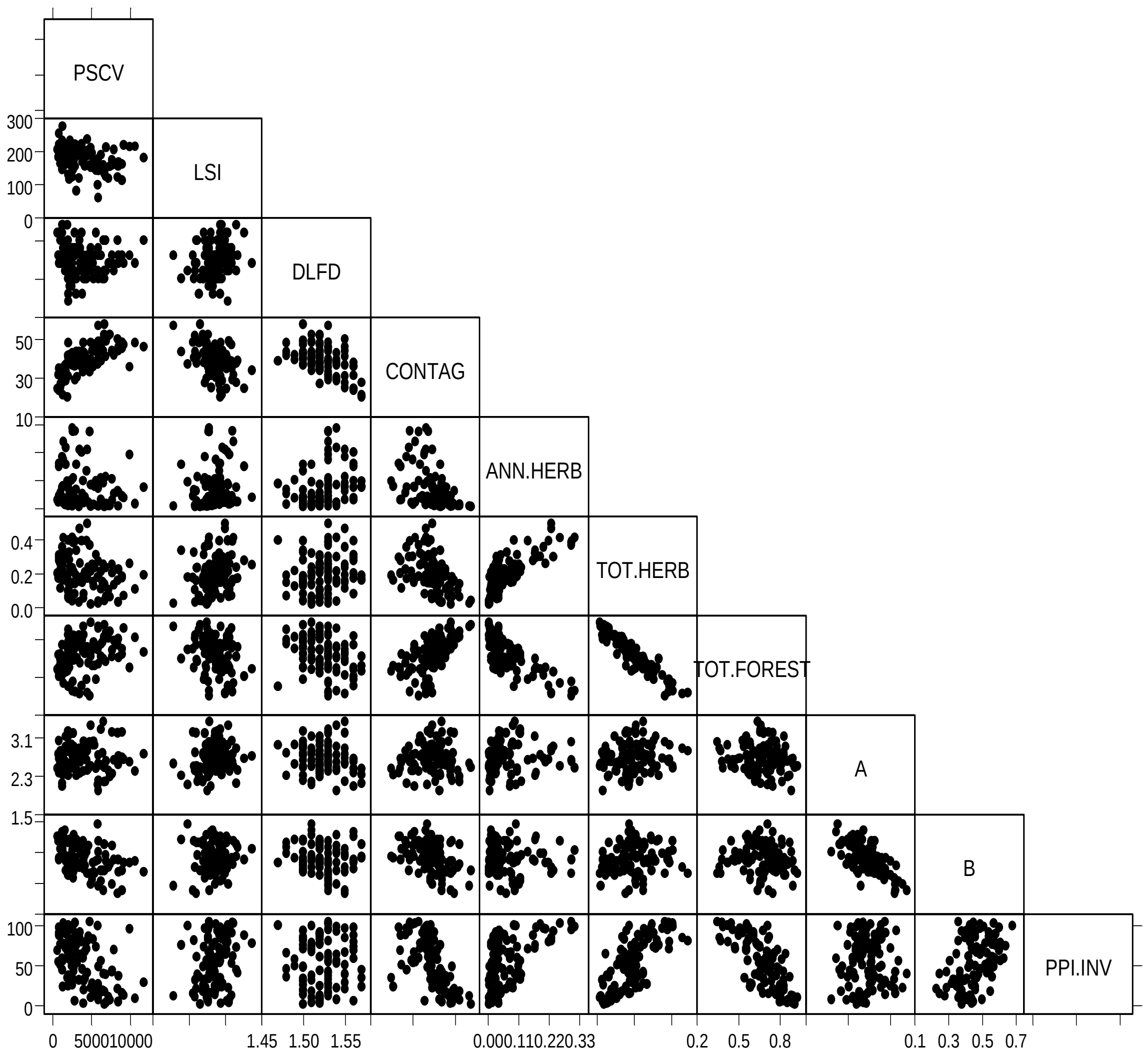
Physiographic stratification indicated by heavy black lines


6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
C
A
B
conditional entropy
0
1
2
3
4
5
6
7
8
9
10
Resolution (power of 2 of unit length of pixel side)
y = C – A*exp(-B*x)
i.e.: Conestoga Creek
A=2.59
B=0.45
C=6.27

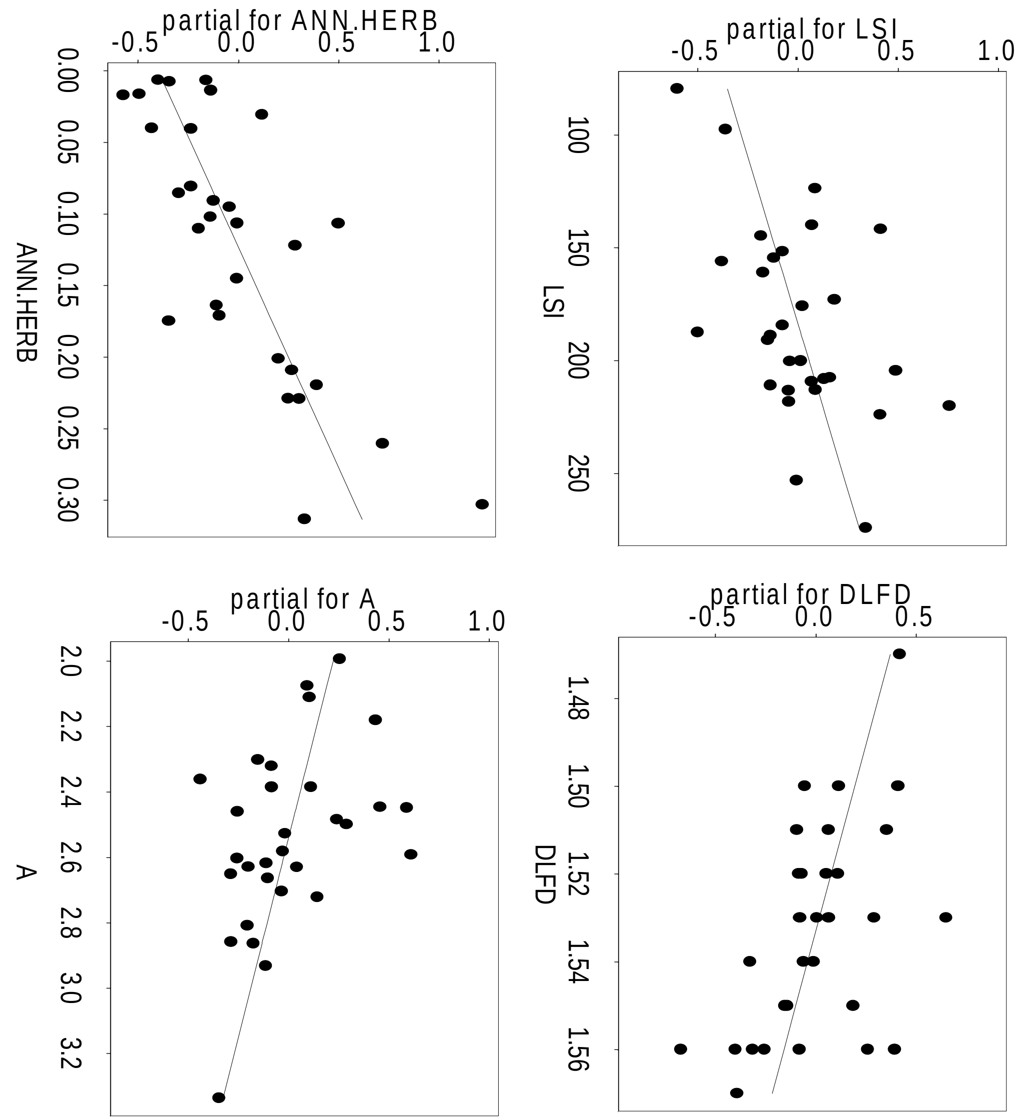


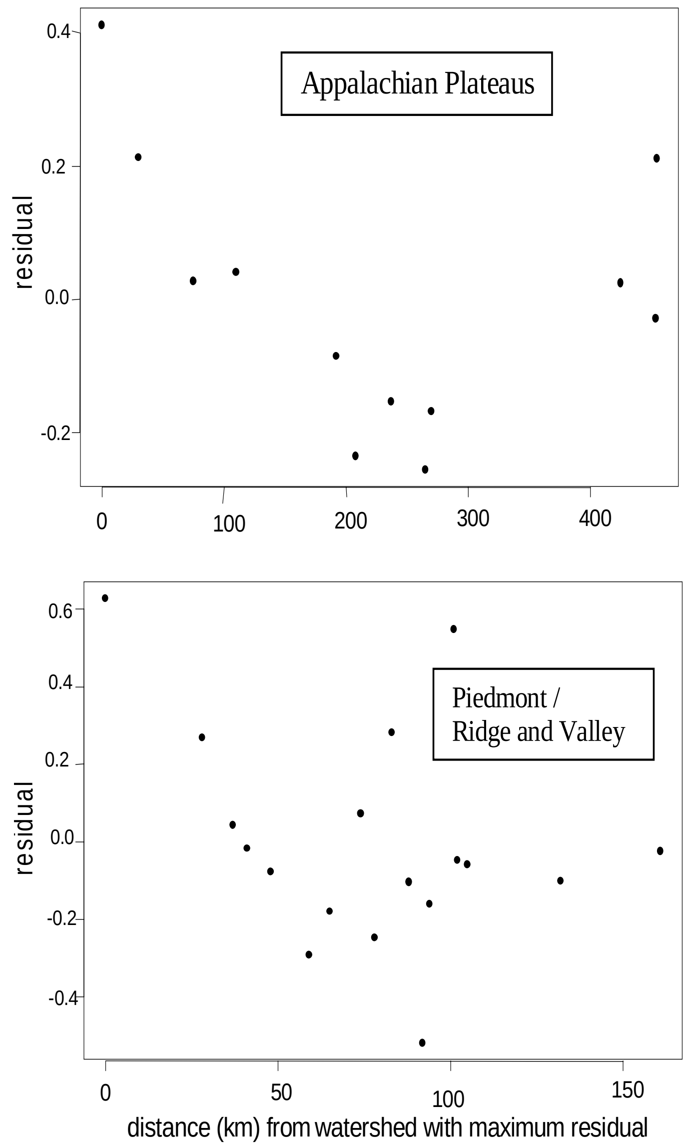
Relative Importance of Principal Components
Variances
0
2
4
6
8
0.628
0.796
0.886
0.932
0.968
0.981
0.989
0.994
0.997
0.999
Comp. 1
Comp. 2
Comp. 3
Comp. 4
Comp. 5
Comp. 6
Comp. 7
Comp. 8
Comp. 9
Comp. 10


PSCV
LSI
DLFD
CONTAG
ANN.HERB
TOT.HERB
TOT.FOREST
A
B
logN
200
100
50
40
30
20
0.4
0.2
0.0
3.1
2.3
1.5
5
4
3
0
5000
10000
1.45
1.50
1.55
0.000.110.220.33
0.2
0.5
0.8
0.2
0.4
0.6

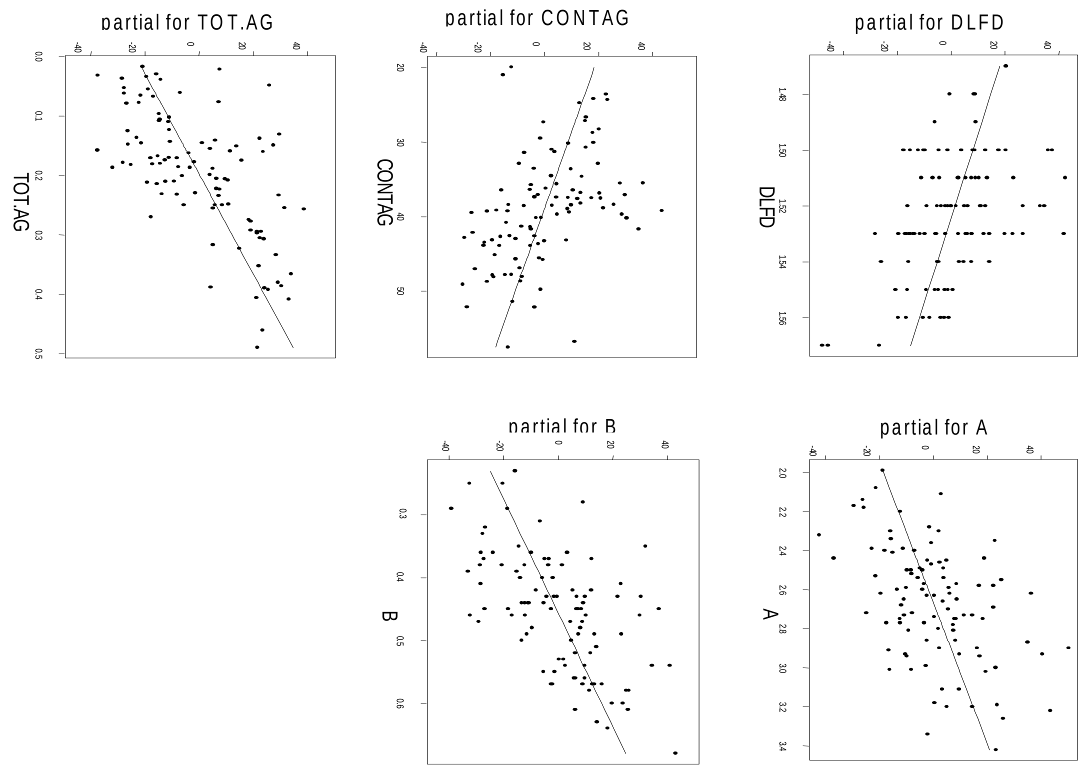


PSCV
LSI
DLFD
CONTAG
ANN.HERB
TOT.HERB
TOT.FOREST
A
B
PPI.INV
300
200
100
0
50
30
10
0.4
0.2
0.0
3.1
2.3
1.5
100
50
0
0 500010000
1.45 1.50 1.55
0.000.110.220.33
0.2 0.5 0.8
0.1 0.3 0.5 0.7

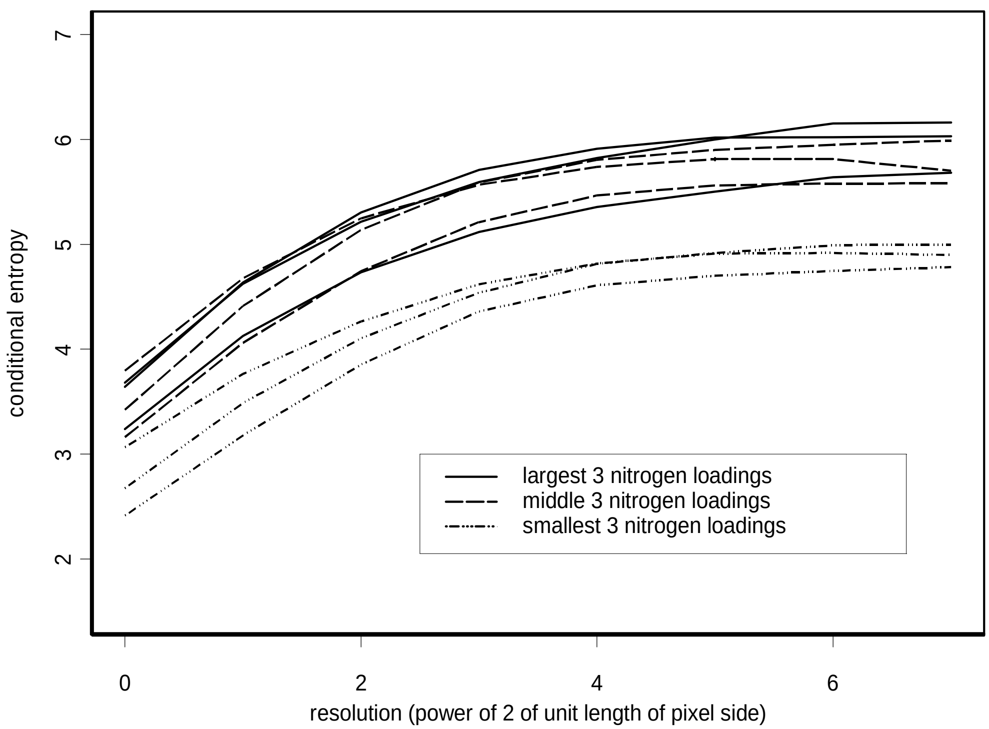

partial for ANN.HERB
-0.5 0.0 0.5 1.0
ANN.HERB
0.00 0.05 0.10 0.15 0.20 0.25 0.30
partial for LSI
-0.5 0.0 0.5 1.0
LSI
100 150 200 250
partial for A
-0.5 0.0 0.5 1.0
A
2.0 2.2 2.4 2.6 2.8 3.0 3.2
partial for DLFD
-0.5 0.0 0.5
DLFD
1.48 1.50 1.52 1.54 1.56



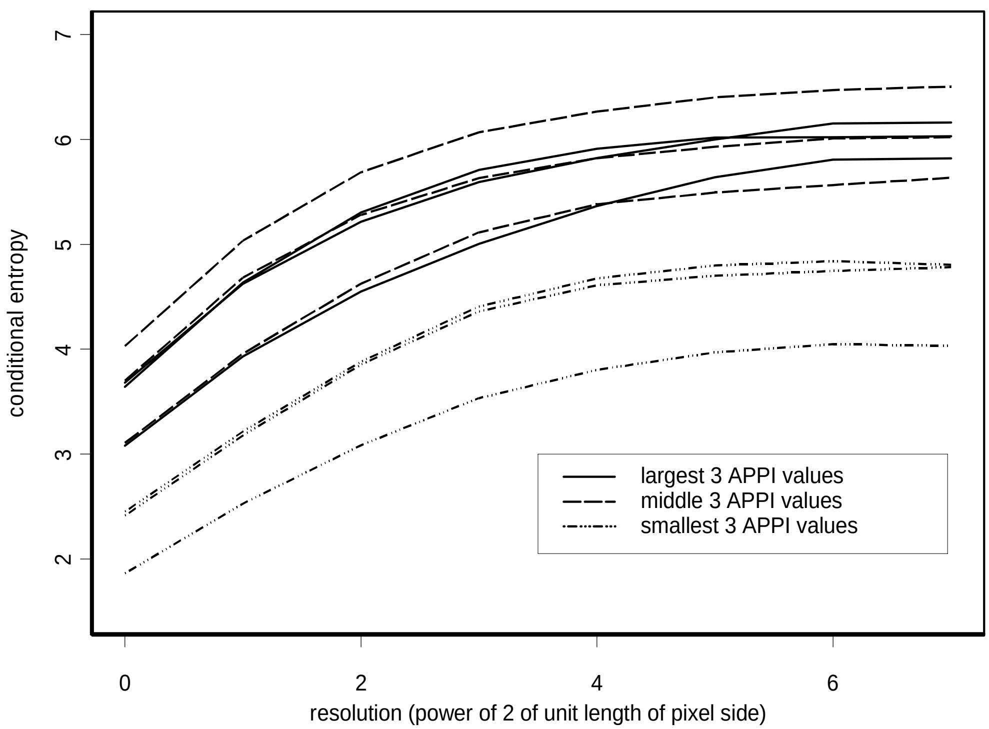
Appalachian Plateaus
residual
0.4
0.2
0.0
-0.2
0
100
200
300
400
Piedmont /
Ridge and Valley
residual
0.6
0.4
0.2
0.0
-0.2
-0.4
0
50
100
150
distance (km) from watershed with maximum residual



partial for TOT.AG
TOT.AG
partial for CONTAG
CONTAG
partial for DLFD
DLFD
partial for B
B
partial for A
A



conditional entropy
7
6
5
4
3
2
0
2
4
6
resolution (power of 2 of unit length of pixel side)
largest 3 nitrogen loadings
middle 3 nitrogen loadings
smallest 3 nitrogen loadings



conditional entropy
2
3
4
5
6
7
0
2
4
6
resolution (power of 2 of unit length of pixel side)
largest 3 APPI values
middle 3 APPI values
smallest 3 APPI values